RSAC | 2025 Conference

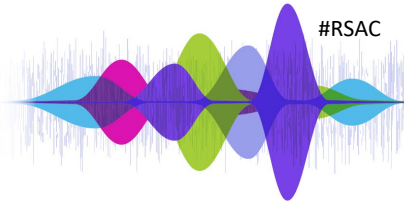
Many Voices.
One Community.

SESSION ID: PDP-M02

Zero Trust AI: Securing Multi-Agent Systems for Private Data Reasoning

Ken Huang

CAIO, Distributedapps.ai
https://www.linkedin.com/in/kenhuang8/



Disclaimer

Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the presenters individually and, unless expressly stated to the contrary, are not the opinion or position of RSA Conference LLC or any other co-sponsors. RSA Conference LLC does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented.

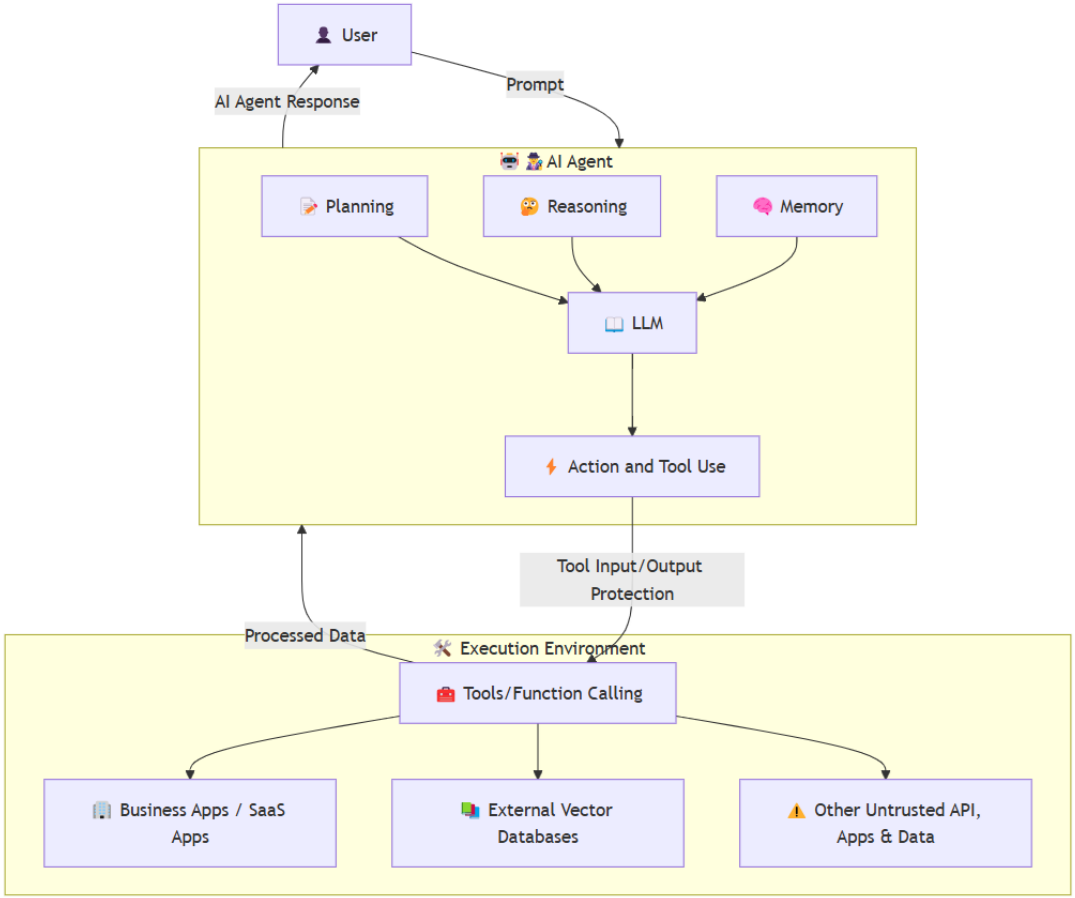
Attendees should note that sessions may be audio- or video-recorded and may be published in various media, including print, audio and video formats without further notice. The presentation template and any media capture are subject to copyright protection.

© 2025 RSA Conference LLC or its affiliates. The RSAC and RSAC CONFERENCE logos and other trademarks are proprietary. All rights reserved.

RSAC 2025 Conference

1: What is Agentic AI?

What is an AI Agent?



No Standard Definition, but we can agree on the **Key Components:**

- **Reasoning** – Logical decision-making
- **Planning** – Strategy & task routing
- **Memory** – Both Short and Long Term
- **LLM** – The brain
- **Autonomy & Tool Use** – Executes tasks

From Single Agent to MAS

Multi-agent systems (MAS) offer several advantages over single-agent systems:

- Enhanced Problem-Solving: MAS outperforms Single Agent
- Scalability: Parallel execution of Agents
- Fault Tolerance: MAS provide redundancy; if one agent fails, others can assume its responsibilities
- Expanded Attack Surfaces: Reduced observability

RSAC | 2025 Conference

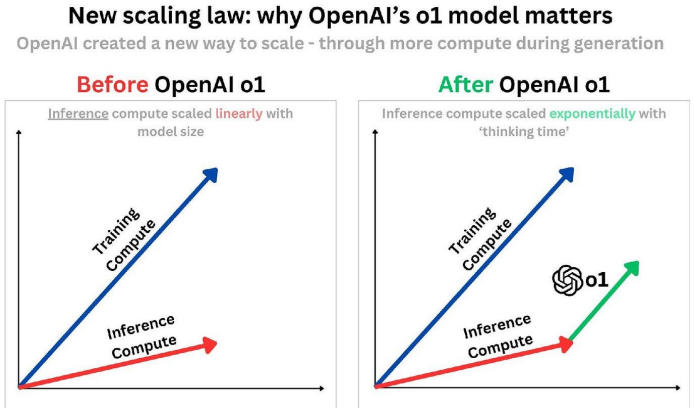
2: Why Everyone IS Talking About Agentic AI?

Key Trends Driving Agentic AI

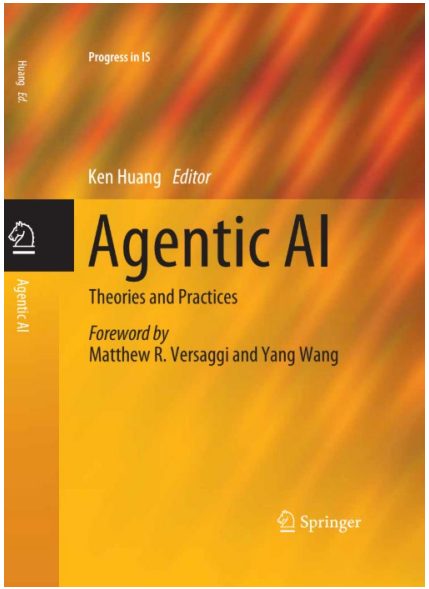
- New Scaling Law
- SaaS is hard
- Physical AI
- No Code, Low Code Tool
- Agentic AI Frameworks
- AI Safety vs. AI Security

AI Trends 1: New Scaling Law

- Scaling on data, computational power and model parameters hits the diminished return
- Scaling on multi-turn reasoning, test time compute, MOE, RL on verifiable domain
- Even if model performance stays at current level, it would take 10 years for enterprise to catch up to mine the intelligence of current model


New scaling law: why OpenAI's o1 model matters
OpenAI created a new way to scale - through more compute during generation
Before OpenAI o1
Inference compute scaled linearly with model size
Training Compute
Inference Compute
After OpenAI o1
Inference compute scaled exponentially with 'thinking time'
Training Compute
o1
Inference Compute

AI Trends 2: End of SaaS and Rise of AI Agents



https://www.amazon.com/Agentic-AI-Theories-Practices-Progress/dp/3031900251

AI Trends 3: Physical AI

- New form factors and computing platforms: Glass? AI Pin? Others?
- Multimodality and New Worldview
- Spatial intelligence (Fei-Fei Li's World Labs)
- Humanoid Robot (10 trillion dollar Company)

https://kenhuangus.substack.com/p/the-new-frontier-securing-physical

AI Trends 4: Code Agent and Vibe Coding

1. Replit AI: AI-assisted coding, instant debugging, refactoring assistance.
2. Cursor AI: AI code completion, smart code improvement, plain language commands.
3. Windsurf AI: Cascade Memories, Cascade Modes, Dead Code Cleanup.
4. Devin AI: Autonomous software development, natural language understanding, comprehensive debugging.
5. Bolt.new: Full-stack development in-browser, AI assistance, npm package integration.
6. Swe-Agent: Specialized debugging, strong IDE integration, error-heavy environment support.

Secure Vibe Coding Guide: https://kenhuangus.substack.com/p/secure-vibe-coding-guide?utm_campaign=post&utm_medium=web

AI Trends 5: Agentic AI Frameworks

Letta: Deep integration with shared memory for persistent, collaborative workflows.

Swarm: Lightweight task handoffs for dynamic, loosely coupled agent collaboration.

AutoGen: Enterprise-grade asynchronous communication for scalable multi-agent systems.

LangGraph: Hierarchical workflows with human-in-the-loop for supervised collaboration.

CrewAI: Role-based collaboration with dynamic task assignment for team-like environments.

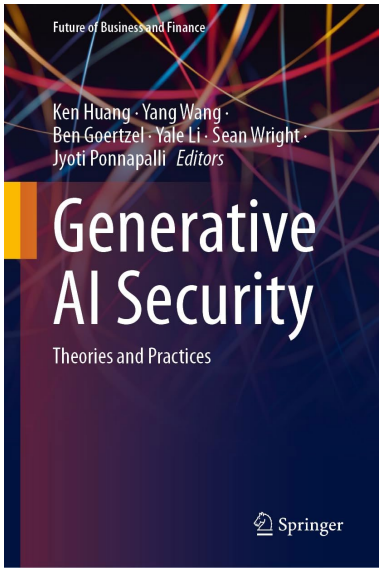
SmolAgents: Ideal for quick prototyping and secure code execution, but limited scalability.

Internal Frameworks: Used by enterprises like Deutsche Telekom (LMOS) and IBM (BEE-AI, open source).

SaaS Frameworks: Solutions such as Agentforce and ServiceNow agents.

AI Trends 6: AI Safety and AI Security

- AI Security focuses on C.I.A
- AI Security focuses on CBRN(Chemical, Biological, Radiological, and Nuclear), Physical Harm, Ethics, AI Doomsday Theory
- AI Security is underhyped, AI Safety is overhyped, which inhibits real world Use of GenAI in Enterprise


Future of Business and Finance
Ken Huang · Yang Wang ·
Ben Goertzel · Yale Li · Sean Wright ·
Jyoti Ponnapalli Editors
Generative AI Security
Theories and Practices
Springer

UK drops 'safety' from its AI body, now called AI Security Institute, inks MOU with Anthropic

https://www.amazon.com/Generative-AI-Security-Theories-Practices/dp/3031542517

https://techcrunch.com/2025/02/13/uk-drops-safety-from-its-ai-body-now-called-ai-security-institute-inks-mou-with-anthropic

RSAC | 2025 Conference

3: What Are Top Agentic AI Threats?

Agentic AI Top Threats

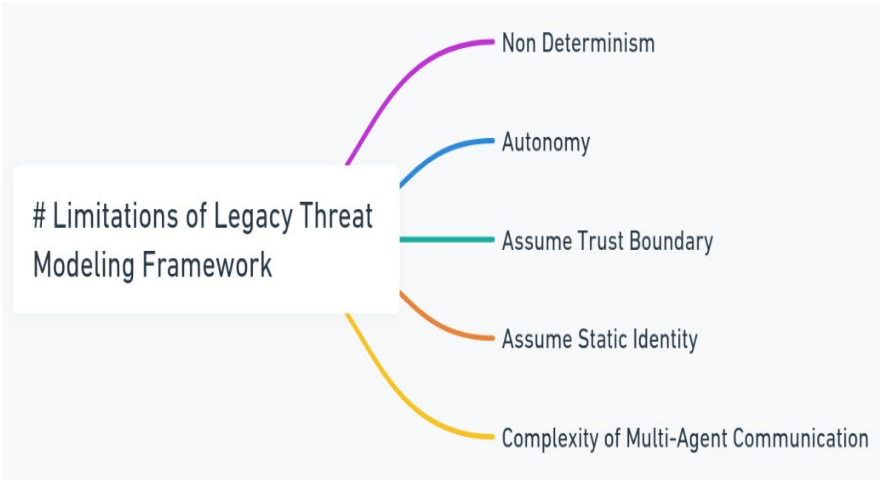
- Agent Authorization and Control Hijacking
- Agent Critical Systems Interaction
- Agent Goal and Instruction Manipulation
- Agent Hallucination Exploitation
- Agent Impact Chain and Blast Radius
- Agent Memory and Context Manipulation
- Agent Orchestration and Multi-Agent Exploitation
- Agent Supply Chain and Dependency Attacks

RSAC | 2025 Conference

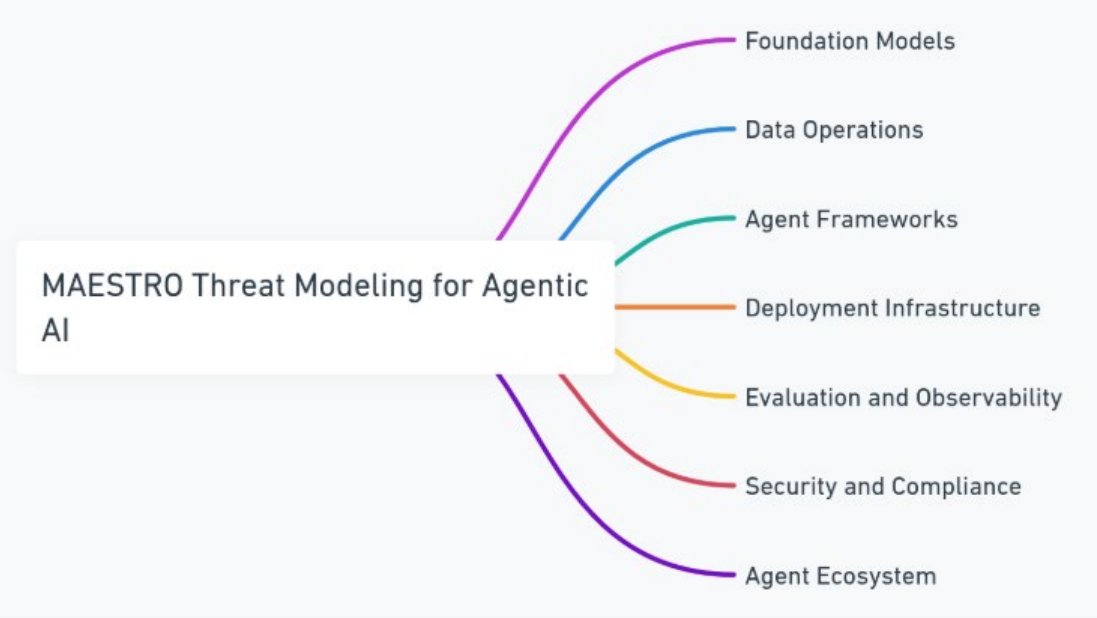
4: Agentic AI Threat Modeling

Why We Need A New Threat Modeling Framework For Agentic AI?

Framework	Focus	Methodology	Use Cases
STRIDE	Six threat categories	Predefined threat categorization	Small teams, app/network security
DREAD	Quantitative risk scoring	Numerical risk prioritization	Risk prioritization, mature orgs
PASTA	Aligns threats with business goals	Risk-centric, SDLC integration	Large enterprises, compliance
OCTAVE	Stakeholder-driven operational risk	Asset-based analysis with collaboration	Structured risk assessment
LINDDUN	Privacy threats	Privacy-focused risk analysis	Healthcare, finance, privacy apps

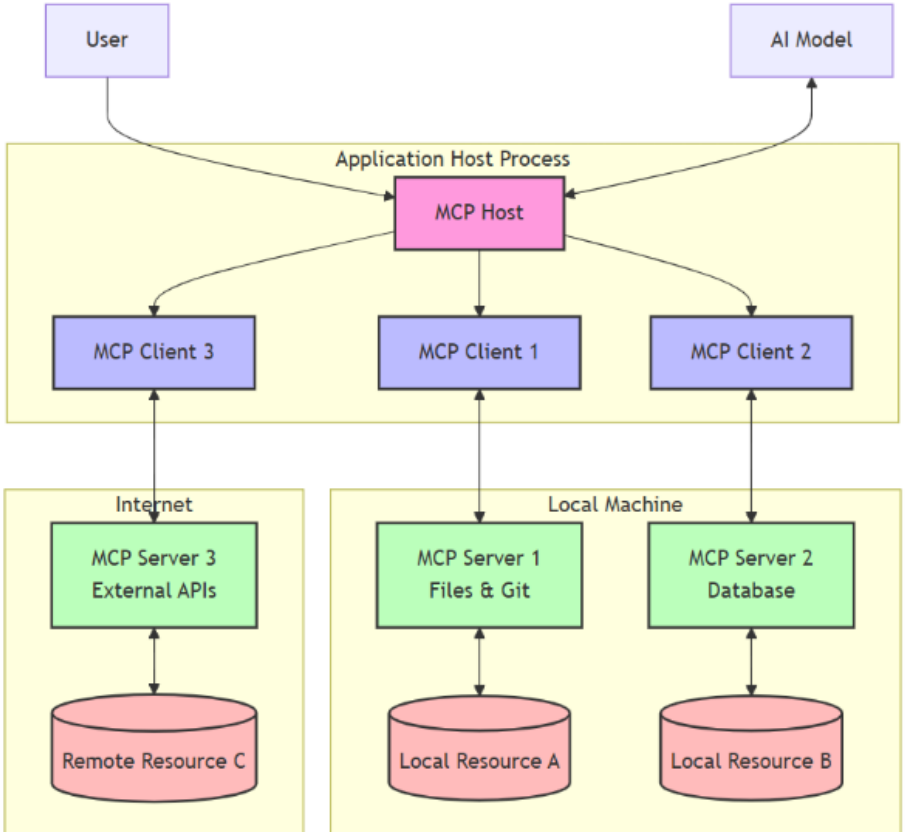


MAESTRO -7 Layer Threat Modeling Approach



Use MAESTRO To Threat Model Apps Built On Top Of MCP

What is the MCP?



How To Perform MAESTRO Threat Modeling (Step 1: Layer Mapping)

MAESTRO Layer	Anthropic MCP Mapping
1. Foundation Models	External LLMs (used by, but *not part of* MCP)
2. Data Operations (RAG)	Servers provide data access (often for RAG)
3. Agent Frameworks	MCP *is* the framework (Host, Client, Protocol)
4. Deployment Infra.	MCP Servers (and their environment)
5. Eval & Observability	Implementation-specific logging/monitoring
6. Security & Compliance	Controlled access, permissions (design principle)
7. Agent Ecosystem	Connects agents to broader ecosystem (tools, data)

How to Perform MAESTRO Threat Modeling (Step 2: Focus On Agentic Factor)

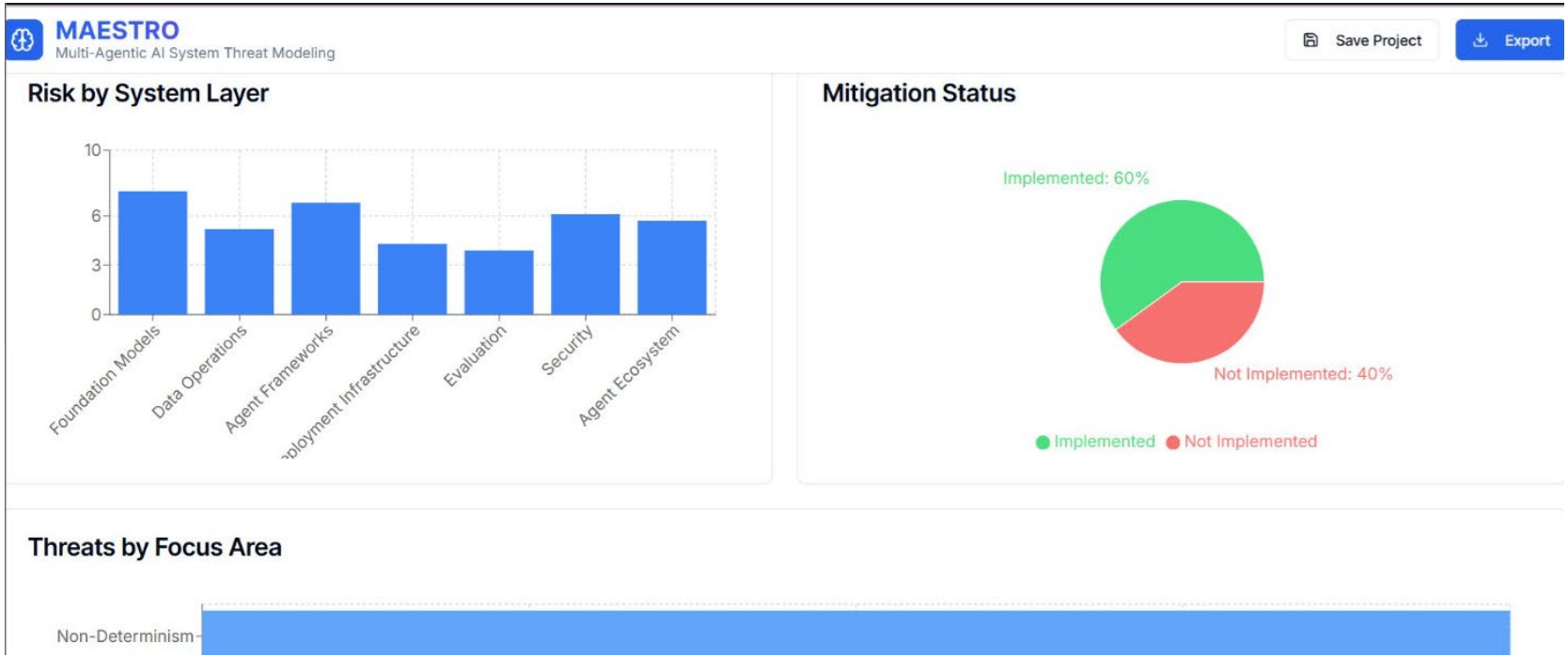
Focus on the following Agentic Factors in Threat Modeling:

- Non determinism
- Autonomy
- No trust boundary
- Dynamic identity and access Control
- Complexity of multi-agent communication and workflow

Sample Threats For MCP

MAESTRO Layer	MCP Threat	Short Description	Driving Factor
1. Foundation Models	Unpredictable API Interactions	LLM may cause unexpected API calls via MCP.	Non-Determinism
2. Data Operations (RAG)	Stale Data, Bad Decisions	Agent uses outdated RAG data, causing errors.	Non-Determinism (Old Info)
3. Agent Frameworks	Runaway Resource Use	Agent overuses MCP, leading to high costs/DoS.	Autonomy
4. Deployment Infra.	Key Rotation Failure	Compromised keys stay valid, risking access.	Dynamic IAM
5. Eval & Observability	Evasion via Mimicry	Malicious agent mimics normal behavior.	No Trust Boundary
6. Security & Compliance	Policy Enforcement Failure	Flawed policy allows incorrect agent actions.	Dynamic IAM
7. Agent Ecosystem	Rogue MCP Server	Fake server impersonates a real one.	No Trust, Agent Identity

MAESTRO Open Source Tool To Manage Threats: Working In Progress



RSAC | 2025 Conference

5: Zero Trust For Agentic AI Systems

Zero Trust: Not a Buzzword, but a Necessity for Agentic AI

- Assume ALL agents are potentially compromised
- Assume ALL responses from Model are untrusted
- Assume ALL external APIs, Vector DB, RAG Pipeline are not secure
- Apply Zero Trust principles at every level: Identity, data, tool use, network, end point
- Continuous verification and validation becomes the norm
- Micro Segmentation of Agent Systems

Agent Identity: Beyond Static Credentials

- Ephemeral Authentication
- Identity Monitoring Agent
- Continuous, real-time policy enforcement

Refer to my CSA' Blog Post:
https://cloudsecurityalliance.org/blog/2025/03/11/agentic-ai-identity-management-approach

Data Security: Beyond "Need to Know" - Context is King

1. Identity Based Security + Data Centric Security

1. Use GenAI to do data labeling

1. Context: Location, time, task-specific permissions

Example: research from West Point
https://www.ieworldconference.org/content/WP2024/Papers/GDRKMCC24_23.pdf

Secure and Verifiable Communication

1. End-to-end encryption

1. Goal alignment check

1. Prevent Agent-in-the-Middle attack

Red Teaming Agentic AI

- Prompt injection: simulating attacks
- Behavioral analysis: detecting anomalies
- Red teaming early and often
- Open source Agent Dojo tool (https://github.com/ethz-spylab/agentdojo)

Data Reasoning without Exposure

- Differential Privacy techniques
  - NIST: Guidelines for Evaluating Differential Privacy Guarantees
- Federated Learning, bring the model to the data
- Data minimization & de-identification

Blockchain: A Decentralized Foundation for Trust

- For high value MAS system
- Immutable audit trails for all AI agent actions and data access
- Decentralized Identity for secure agent ID management

RSAC 2025 Conference

6: Case Studies And How To Apply What You've Learned

Case Study 1: Open Source Zero Trust(ZT) SDK for Agentic Frameworks

- Authentication: Securely verify agent identities.
- Policy Enforcement: Fine-grained control of agent actions.
- Security Monitoring: Tracks and audits security events.
- Framework Adapters: Integrates with CrewAI and AutoGen.
- Configuration: .env and policy.yaml for settings.
- OAuth & Social Login: Supports Google, GitHub, Entra ID.

```python
from zta_agent import initialize_agent
from crewai import Agent, Task, Crew

zta = initialize_agent()
auth_manager, crewai_adapter = zta["auth_manager"], zta["crewai_adapter"]


def authenticate_and_validate(identity, secret):
    token = auth_manager.authenticate({"identity": identity, "secret": secret})
    return crewai_adapter.validate_agent_action(
        identity, {"type": "execute_task", "resource": "operation"}, token
    )


if authenticate_and_validate("research_agent", "research_secret"):
    agent = Agent(
        role="Researcher",
        goal="Research",
        backstory="...",
        allow_delegation=False,
    )
    task = Task(description="Research AI security patterns", agent=agent)
    result = Crew(agents=[agent], tasks=[task]).kickoff()
    print("Crew execution result:", result)
else:
    print("Execution denied.")
```

Github: https://github.com/kenhuangus/ZeroTrustAgent

Case Study 2: Agent Monitoring and Enforcement

Stealth startup

- Real-time monitoring of AI agents
- Visibility into agents' actions
- Enforcement of policies
- AI-supervision for AI

Apply Today: Actionable Steps

1. Threat Model Your Agents Using a Layered Approach: https://cloudsecurityalliance.org/blog/2025/02/06/agentic-ai-threat-modeling-framework-maestro

1. Implement ZT Principles and Explore and utilize the open-source SDK: https://github.com/kenhuangus/ZeroTrustAgent

1. Red Teaming Your Agent

1. Continuous Monitoring

Key Takeaways

1. What, Why, How of Agentic AI and Security

1. Multi-Agent Systems (MAS) are powerful but risky

1. Agentic AI has unique security threats, Zero Trust is needed

1. Red Teaming early and Red Teaming often

RSAC | 2025 Conference

Many Voices.
One Community.

Thank you

For more information, connect with me via LinkedIn
https://www.linkedin.com/in/kenhuang8/